Global Scientific JOURNALS

GSJ: Volume 9, Issue 7, July 2021, Online: ISSN 2320-9186
www.globalscientificjournal.com

PRINCIPAL COMPONENT ANALYSIS (PCA)

(Using Eigen Decomposition)

Aatif Nisar Dar AATIF.DAR11@GMAIL.COM

Department of Computer Science

South Asian University

Akbar Bhawan, Chanakyapuri, New Delhi 110021, India

© GSJ

Abstract:

This paper will discuss Principal Component Analysis (PCA), which is used to reduce the dimensionality of a dataset. We achieve this reduction of dimensionality by transforming this dataset to a new dataset of uncorrelated principal components or variables, or features. PCA is a multivariate technique, and the Principal components are the Eigenvectors of the new data's covariance matrix. PCA is a potent tool for analyzing the data by finding the patterns in the data and reducing the number of dimensions without much loss of information. PCA is used in many applications like multivariate data analysis, image compression, face recognition, and many more.

Introduction:

Principal Component Analysis was first introduced by Pearson in 1901 and developed independently by Hotelling in 1933. Pearson states that his methods can be easily applied to numerical problems. Although he says that the calculations become challenging to carry for four or more variables, he suggests that they are still quite feasible. It is not possible to do PCA by hand unless we have four variables or fewer. But it is precisely for variables greater than four that PCA is most beneficial. To analyze the data by Principal Component Analysis, we have to be thorough in statistics and matrix algebra. The central idea of PCA is to identify correlations and patterns in a dataset of higher dimensions and reduce it to a significantly lower dimension without loss of any vital information. The need for the PCA technique is because the high dimensionality data is highly complex due to inconsistencies in the features that increase the computation time. Principal Components are given by an orthogonal linear transformation of a set of variables optimizing a specific algebraic criterion.

PCA is an unsupervised method in that it does not use the output information; the criterion to be maximized is the variance.

Keywords: Covariance matrix; Eigenvalues; and Eigenvectors; Projection; Multivariate; Variance.

Steps involved in the PCA algorithm are as follows:

1) Data Scaling
2) Computing the covariance matrix
3) Calculating Eigenvalues and Eigenvectors
4) Computing the principal components
5) Deriving the new dataset

Mean: We sum up all data points in our data set and divide by the number of data points that we have. Mean represents an average data point. Mean is also known as the expected value of a dataset.

$$\boldsymbol{E(X) = X` = \frac{1}{N} * \sum_{i=1}^{N} x_i}$$

Where,

X` is mean

N is data points

$E(X)$ = Expected value

$\sum_{i=1}^{N} x_i$ is the sum of data points

© GSJ

Variance: The variance is used to characterize the spread of data points in a dataset. In one dimension, we can look at the average squared distance of a data point from the mean value of this dataset. Variance is denoted by σ^2.

$$\boldsymbol{Var(X) = \frac{1}{N-1} * \sum_{i=1}^{N} (x_i - X')^2}$$

In higher dimensional datasets, we check if there is a relationship between any dimensions. For this purpose, we use **Covariance**. Covariance is a similarity measure to find out how much the dimensions vary from the mean with respect to each other.

$$\boldsymbol{Cov(X,Y) = \frac{1}{N-1} * \sum_{i=1}^{N} (x_i - X`)(y_i - Y`)}$$

Where,

X`, Y` is the mean of datasets X and Y, respectively.

N is the number of observations.

If the Covariance between X and Y is positive, then on average, the Y value increases if we increase X. And if the Covariance between X and Y is negative, then the Y value decreases if we increase X on

average. If the Covariance between X and Y is zero, X and Y have nothing to do with each other. They are uncorrelated.

Standard deviation: is a measure of the amount of variation or dispersion of a set of values. The standard deviation of a set of observations of a series is the positive square root of the arithmetic mean of the squares of all the deviations from the arithmetic mean. Standard deviation is denoted by σ.

$$\boldsymbol{S.D} = \sqrt{\frac{1}{N-1} * \sum_{i=1}^{N} (x_i - X`)^2}$$

Covariance Matrix: is a square matrix that expresses the correlation between different variables. For the n dimension dataset, the size of our covariance matrix becomes n*n, and each entry is the matrix is the result of calculating the Covariance between two separate dimensions. For the n dimension dataset, we calculate

$\frac{n!}{(n-2)*2}$ different covariance values. The covariance matrix defines both the spread (variance) and the orientation (Covariance) of our data.

Properties of covariance matrix:

1) The Covariance matrix is a Squared matrix
2) Covariance matrix is a symmetric matrix i.e. $cov(X,Y) = cov(Y,X)$
3) Covariance matrix is positive semidefinite matrix i.e. $A\Sigma A^T >= 0$

Covariance matrix for a 'd' dimensional dataset:

Diagonal elements are the covariance value between one of the dimensions with itself. We write variance in place of Covariance on diagonal elements as $cov(X,X) = var(X)$.

$$\Sigma = \begin{bmatrix} var(x_1,x_1) & \cdots & cov(x_1,x_d) \\ \vdots & \ddots & \vdots \\ cov(x_d,x_1) & \cdots & var(x_d,x_d) \end{bmatrix}$$

Dot Product:

The dot product between two vectors, x, and y is defined as:

$$\boldsymbol{x^T.y = \sum_{i=1}^{N} x_i.y_i \ , \ x,y \in Rn}$$

Two vectors are orthogonal if they are perpendicular. In mathematical notation:

$$\boldsymbol{x^T.y = 0}$$

Length of x is defined as:

$$||x|| = \sqrt{(x^T.x)} = \sqrt{\sum_{i=1}^{N} x_i{}^2}$$

Distance between two vectors x and y is defined as:

$$\boldsymbol{d(x,y) = ||x-y|| = \sqrt{(x-y)^T(x-y)}}$$

The angle between vector x and y is defined as:

$$\boldsymbol{cos\ \alpha = \frac{x^T.y}{||x||\ ||y||}}$$

Inner Product:

An inner product is a generalization of the dot product. In a vector space, it is a way to multiply vectors together, with the result of this multiplication being a scalar.

Let V be a vector space over R. An inner product (,) is a function V × V → R with the following properties:

1) Symmetric $(\boldsymbol{u,v}) = (\boldsymbol{v,u})$
2) Positive semidefinite $(\boldsymbol{u,u}) >= 0$
3) Bilinear $(\boldsymbol{\alpha(u,v),w}) = \boldsymbol{\alpha(u,v)} + (\boldsymbol{v,w})$

$$\forall\ \boldsymbol{u,v,w}\ \in\ \boldsymbol{V}$$

Length or Norm function via the inner product is defined as:

$$\boldsymbol{||u|| = \sqrt{(u,u)}}$$

Distance via the inner product is defined as:

$$\boldsymbol{dist(u,v)}\ =\ \boldsymbol{||u-v|| = \sqrt{(u-v,u-v)}}$$

Angle via the inner product is defined as:

$$\boldsymbol{Cos\ \alpha = \frac{(u,v)}{||u||||v||}}$$

Two vectors $u, v\ \in\ V$ are orthogonal, if and only if,

$$\boldsymbol{(u,v) = 0}$$

Orthogonality is defined with respect to the inner product. And vectors that are orthogonal with respect to one inner product do not have to be orthogonal with respect to another inner product.

Eigenvalues and Eigenvectors:

PCA uses Eigen decomposition of the covariance matrix in order to determine the principal components. Eigenvalues and Eigenvectors exist in pairs, i.e., every Eigenvector has a corresponding eigenvalue. For an n*n covariance matrix, we will have n Eigenvectors. Eigenvectors are used to understand variance (spread) in our dataset, i.e., in which variable we have more variance and the Eigenvector will be equal to the magnitude of that direction. If we sort the Eigenvalues in descending order, the Eigenvector associated with the first Eigenvalue gives us the first principal component (PC1), the second Eigenvector associated with the second Eigenvalue gives us the second principal component (PC2), and so on.

Some properties of these eigenvectors:

- Eigenvectors of the covariance matrix are always orthogonal (perpendicular) to each other, and the data is expressed in terms of these orthogonal Eigenvectors.
- When a linear transformation (multiply them with another vector) is performed on them, their direction does not change.
- We are only concerned with the direction of the vector and not the length. Hence the length of the Eigenvectors is set to 1 so that all eigenvectors will have the same length.
- Eigenvector of the covariance matrix $\sum$ satisfies the following equation-

$$\sum \boldsymbol{u} = \boldsymbol{\lambda u}$$

Where,

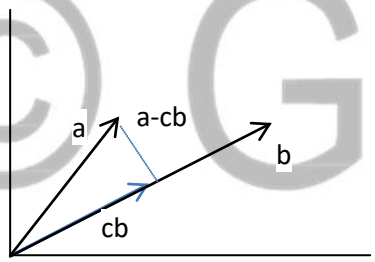
$\boldsymbol{u}$ is Eigenvector

$\boldsymbol{\lambda}$ is Eigenvalue corresponding to that Eigenvector

PROJECTION:

The first principal component is the unique vector that maximizes the variance of the projection of the data onto that vector. Projection is an operator on two vectors.

Projection of a vector 'a' on a vector 'b' is the vector in the direction of vector 'b' and lets call that 'cb' where 'c' is scaler such that 'a-cb' is orthogonal to vector 'b'. 'a-cb' is called the orthogonal projection of 'a' on 'cb'. It's best understood graphically.



© GSJ

Since 'a-cb' is orthogonal to vector 'b', their inner product is zero i.e.

$$(\boldsymbol{a} - \boldsymbol{cb}, \boldsymbol{b}) = \boldsymbol{0}$$

$$(\boldsymbol{a}, \boldsymbol{b}) - (\boldsymbol{cb}, \boldsymbol{b}) = \boldsymbol{0}$$

$$(\boldsymbol{a}, \boldsymbol{b}) - \boldsymbol{c}(\boldsymbol{b}, \boldsymbol{b}) = \boldsymbol{0}$$

As, $(\boldsymbol{b}, \boldsymbol{b}) = \boldsymbol{b^T b} = \left|\left|\boldsymbol{b}\right|\right|^2$ when we take the inner product as dot product

$$\boldsymbol{c} = \frac{(\boldsymbol{a}, \boldsymbol{b})}{\left|\left|\boldsymbol{b}\right|\right|^2}$$

Where $\left|\left|\boldsymbol{b}\right|\right|^2$ is the squared norm of vector b.

Choosing the inner product to be the dot product

$$\boldsymbol{c} = \frac{\boldsymbol{b^T a}}{\left|\left|\boldsymbol{b}\right|\right|^2}$$

If $||b|| = 1$, then the coordinate 'c' of the projection = $\boldsymbol{b^T a}$

So our projection of vector 'a' on vector 'b' = **cb = bc**

$$= \frac{\boldsymbol{b}\,(\boldsymbol{a},\boldsymbol{b})}{\boldsymbol{||b||2}}$$

$$= \frac{\boldsymbol{bb^T a}}{\boldsymbol{||b||^2}}$$

We know that a projection is a linear mapping .Therefore, there exists a projection matrix P_c such that $cb = P_c * a$

$$\boldsymbol{cb} = \frac{\boldsymbol{bb^T}}{\boldsymbol{||b||^2}} * \boldsymbol{a}$$

Therefore,

$$\boldsymbol{p_c} = \frac{\boldsymbol{bb^T}}{\boldsymbol{||b||^2}}$$

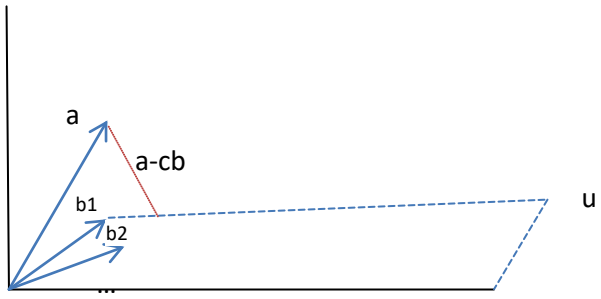
Where,

$\boldsymbol{bb^T}$ is a symmetric matrix and $\boldsymbol{||b||^2}$ is a scaler.

Projection onto higher-dimensional subspaces

First, we will look at the definition of vector basis. A vector basis of a vector space is defined as a subset of vectors that are linearly independent.

Let B = {b1,b2,…} be basis vector. Let 'u' be the plane spanned by basis vector. Let 'a - cB' denote the orthogonal projection of vector 'a' on 'cb' where 'cb' is the projection of vector 'a' in the direction of vector basis B.



The difference of vector 'a' with the projection of vector 'a' on vector 'b' is orthogonal to vector 'b' i.e. its inner product is zero

$$(\boldsymbol{a} - \boldsymbol{cb_i}, \boldsymbol{b_i}) = \boldsymbol{0}$$

Where,

b_i = {b1,b2,…}

$$(\boldsymbol{a}, \boldsymbol{b_i}) - (\boldsymbol{cb_i}, \boldsymbol{b_i}) = \boldsymbol{0}$$

$$\boldsymbol{a^T b_i - c^T b_i{}^T b_i = 0}$$

$$\boldsymbol{a^T B - c^T B^T B = 0}$$

$$\boldsymbol{c^T = a^T B (B^T B)^{-1}}$$

$$\boldsymbol{c = (B^T B)^{-1} B^T a}$$

Matrix $\boldsymbol{(B^T B)^{-1} B^T}$ is also called the pseudo-inverse of B

Projection of vector 'a' on vector 'b' = $\boldsymbol{c * B}$

$$\boldsymbol{= (B^T B)^{-1} B^T a * B}$$

$$\boldsymbol{= B (B^T B)^{-1} B^T a}$$

Projection matrix pc solves $\boldsymbol{p_c * a = cB}$

Therefore, $\boldsymbol{p_c = B (B^T B)^{-1} B^T}$

$$\boldsymbol{p_c = \frac{B B^T}{B^T B}}$$

We just looked at projections of vectors a onto a subspace U with basis vectors {b1, b2,. . . ,}. If this basis is an orthonormal basis vector, the projection equation simplifies to:

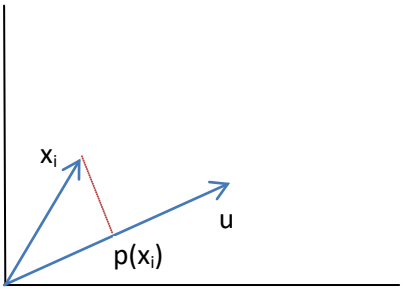
$$\boldsymbol{cB = B B^T a}$$

Since $\boldsymbol{B^T B = I}$

© GSJ

Objectives of PCA Algorithm:

1) Maximize the variance



Let $x_1, x_2, \ldots, x_N$ be N data points in a k- dimensional data space, i.e., $x_i \in R^k$. The N datapoints are mean-centered.

Projection of x_i on u is given by:

$$\boldsymbol{p(x_i) = \left(\frac{uu^T}{u^Tu}\right)x_i}$$

$$\boldsymbol{p(x_i) = \left(\frac{u^Tx_i}{u^Tu}\right)u}$$

$$\mathbf{p(x_i) = (u^Tx_i)u}$$

Because u is a unit vector i.e. $||u|| = 1$ implies $\boldsymbol{u^Tu = 1}$

$$\boldsymbol{p(x_i) = a_iu}$$

where $\boldsymbol{a_i = u^Tx_i}$

Let x` denote the mean of N datapoints i.e. $\boldsymbol{x` = mean\{x_i\} \quad i = 1 \ldots N}$

We seek to find the unit vector u (optimal basis vector) that maximizes the projected variance of the datapoints. $u \in R^k$

Maximum variance along u is given by:

$$\boldsymbol{\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(a_i - x`)^2}$$

$$\boldsymbol{\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}\left(u^Tx_i - u^Tx`\right)^2}$$

$$\boldsymbol{\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}\left(u^Tx_i\right)^2}$$

As the data is mean centered. Hence $x` = 0$

$$\boldsymbol{\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}u^T\left(x_ix_i^T\right)u}$$

$$\boldsymbol{\sigma^2 = u^T\frac{1}{N}\sum_{k=1}^{N}\left(x_ix_i^T\right)u}$$

$$\boldsymbol{\sigma^2 = u^T\sum u}$$

Where $\sum$ is a k*k covariance matrix

Our optimization problem becomes:

Max $\boldsymbol{\sigma^2 = u^T\sum u}$

s.t.

$||\boldsymbol{u}|| = \mathbf{1}$ **implies** $\boldsymbol{u^Tu = 1}$

Where $\boldsymbol{u \in R^k}$

Using Lagrangian multiplier μ

$$\boldsymbol{\sigma^2 = u^T\sum u - \mu\left(1 - u^Tu\right)}$$

Partial differentiating w.r.t. u and making it equal to zero

$$\frac{d(\sigma^2)}{d(u)} = 0$$

$$\frac{d\left(u^T \sum u - \mu(1 - u^T u)\right)}{d(u)} = 0$$

$$2\sum u - 2\,\mu u = 0$$

$$\sum u = \mu u$$

Thus,

$\sum$ = Covariance matrix

u = Eigen Vector

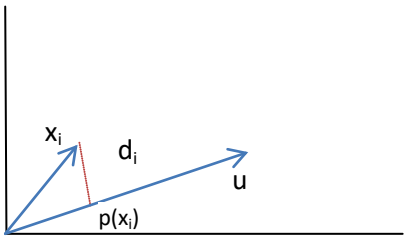
μ = Eigenvalue associated the Eigenvector u of the covariance matrix $\sum$

We maximize projected variance σ^2, we thus choose the Largest Eigenvalue of the covariance matrix, and the corresponding Eigenvalue specifies the direction of the most variance, which is called as First Principal Component.

2) Distance Minimization

The direction that maximizes the projected variance is also the one that minimizes the distance. We seek to find the optimal basis unit vector u that minimizes the squared distance i.e.

$$min \sum_{i=1}^{N} d_i^{\,2}$$



Let di be the distance between x_i and the optimal basis vector u.

We know that,

$$||x_i||^2 = p(x_i) + d_i^{\,2}$$

$$d_i^{\,2} = ||x_i||^2 - u^T x_i$$

$$d_i^{\,2} = (x_i^T x_i - u^T x_i)$$

Our optimization Problem becomes:

$$Min \sum_{i=1}^{N} (x_i{}^T x_i - u^T x_i)^2$$

s.t.

$||u|| = 1$ **implies** $u^T u = 1$

Methodology of Principal Component Analysis Algorithm:

We will understand PCA by analysing a own designed dataset.

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6

1) Data Scaling

Here, N = 7

Now we find the mean of both the features and subtract the mean from each of the data dimension

X` = (2.5 + 0.5 + 2.2 + 1.9 + 3.1 + 2.3 + 2)/7
= 14.5/7
= 2.0714

Y` = (2.4 + 0.7 + 2.9 + 2.2 + 3.0 + 2.7 + 1.6)/7
= 15.5/7
= 2.2142

X = Deviation of x from mean X`	Y = Deviation of y from mean Y`
0.4286	0.1858
-1.5714	-1.5142
0.1286	0.6858
-0.1714	-0.0142
1.0286	0.7858
0.2286	0.4858
-0.0714	-0.6142

2) Computing the covariance matrix

Now we will calculate the covariance matrix of the scaled data.

```
Cov= [0.63571429, 0.58547619
      0.58547619, 0.67142857]
```

Since the data is 2 dimensional, the size of covariance matrix will be 2×2. Also, both the features increase together as the non-diagonal elements in this covariance matrix are positive.

3) Calculating Eigenvalues and Eigenvectors

Since the covariance matrix is square, the eigenvectors and eigenvalues for the matrix can be calculated.

Eigenvectors of the covariance matrix are:

```
[[-0.7178043, -0.69624492],
       [0.69624492, -0.7178043]]
```

Eigenvalues associated with the above Eigenvalues of the covariance matrix are:

```
[0.06782298, 1.23931988]
```

4) Computing the principal components

The eigenvector with the highest eigenvalue is the principle component of the data set.

1.23931988 > 0.06782298.

Hence Eigenvector **[0.69624492, -0.7178043]** which is associated with the Eigenvalue **1.23931988** will give us the First Principal Component.

We will explain this with the concept of Explained Variance. The explained variance ratio is the percentage of variance that is attributed by each of the selected components.

The explained variance of the two vectors is:

```
[0.948113568624776, 0.05188643137522321]
```

The sum of explained variances is always equal to 1. The first two principal components account for around 94% of the variance in the dataset.

5) Deriving the new dataset

Now we derive our new dataset from the chosen components (Eigenvectors).

Final Data = Row Feature Vector × Row Data Adjust

Where,

Row Feature vector = matrix with the eigenvectors in the rows

Row Data Adjust = matrix with data items are in each column, with each row holding a separate dimension.

PC (Principal component)
-0.178289
0.073704
0.385175
0.113145
-0.191224
0.174146
-0.376382

As expected, it only has a single dimension. Therefore, we have successfully reduced our two dimensional dataset to one dimension.

Conclusion:

The benefit of Principal Component Analysis is that we can find the larger variances associated with the first Principal Components and then a precipitous drop up. PCA is a useful statistical method to abstract special features from a data set with a high number of attributes. PCA is a non-parametric analysis technique which is its both advantage and disadvantage. Performing PCA is quite simple. The appropriate way is to first convert the data to the appropriately centered polar coordinates and then compute PCA. We organize a data set as an m×n matrix;

Where m = number of measurement types or row

n = is the number of trials or column

Subtract the mean from each row. Calculate the eigenvectors of the Covariance. Compute the Principal Components using the eigenvectors and extract the new dataset. There are many interesting applications of PCA, out of which in today's life knowingly or unknowingly, multivariate data analysis and image compression are being used alternatively.

References:

- Principal Component Analysis, Second Edition - IT Jolliffe
- Introduction to Machine Learning Second Edition - KK Patel
- Pattern Recognition and Machine Learning - Christopher M. Bishop
- Mathematics for Machine Learning - Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong
- Principal_Component_Analysis - Sidharth Mishra, Uttam Sarkar, Subhash Taraphder, Sanjoy Datta
- Data Mining and Machine Learning - Zaki & Meira Jr.
- Inner Product -- from Wolfram MathWorld. https://mathworld.wolfram.com/InnerProduct.html
- Summary of Video - Learner. https://www.learner.org/wp-content/uploads/2019/03/AgainstAllOdds_StudentGuide_Unit06-Standard-Deviation.pdf
- Variance of one-dimensional datasets - Statistics of https://www.coursera.org/lecture/pca-machine-learning/variance-of-one-dimensional-datasets-1i0Li

- Variance of higher-dimensional datasets - Statistics of .... https://www.coursera.org/lecture/pca-machine-learning/variance-of-higher-dimensional-datasets-QCWpn
- Urban-rural differentials in the association between HIV .... https://iussp.org/sites/default/files/event_call_for_papers/urban%20-rural%20differentials%20in%20the%20link%20between%20HIV%20infection%20and%20poverty%20in%20Kenya%20-%20draft%20(version%202).pdf

© GSJ